IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants:	Pablo Tamayo, Jill Mesirov, Eric S. Lander, Todd R. Golub		
Application No.:	09/525,142	Group:	1631
Filed:	March 14, 2000	Examiner:	S. Zhou

For: METHODS AND APPARATUS FOR ANALYZING GENE EXPRESSION DATA

#15
Plunkett
2/7/02

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as First Class Mail in an envelope addressed to Assistant Commissioner for Patents, Washington, D.C. 20231

on 1/28/02 Annie Demirel
Date Signature

Annie Demirel
Typed or printed name of person signing certificate

DECLARATION OF PABLO TAMAYO, JILL MESIROV, ERIC S. LANDER, AND TODD R. GOLUB UNDER 37 C.F.R. § 1.131

Assistant Commissioner for Patents
Washington, D.C. 20231

We, Pablo Tamayo, Jill Mesirov, Eric S. Lander, and Todd R. Golub declare as follows:

1. We are the co-inventors of the above-identified U.S. Patent Application.

2. We have read U.S. patent application 09/525,142 and the Office Actions mailed on May 7, 2001 and October 26, 2001.

3. We hereby state that we reduced to practice the claimed invention for the above-referenced application prior to December 1998, the effective publication date of Eisen, M. B., *et al.*, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *PNAS*, 95:14869-14868 (1998) (hereinafter "Eisen").

4. Evidence of actual reduction to practice of the invention is presented by the attached draft manuscript, entitled, "Self-Organized Maps of Gene Expression: Applications to Hematopoietic Differentiation and the Cell Cycle," (hereinafter "Exhibit A") that was prepared and completed by us prior to December 1998. The manuscript was prepared for submittal for publication in the journal, Nature Genetics. Exhibit A describes use of Self-Organizing Maps (SOM) to classify and analyze gene expression data, and contains the data that served as examples in the above-referenced patent application. The software that employs the methodology of the claimed invention was written and tested prior to December 1998. Exhibit A states that the claimed methodology was tested on gene expression data relating to hematopoietic differentiation (pages 2-5), and the yeast mitotic cell cycle (pages 5-6). Exhibit A further states that SOM is an effective tool for classifying gene expression data. Exhibit B is additional evidence that the draft manuscript was written and ready for publication at least by October 23, 1998.

5. We further declare that all statements made herein of our own knowledge are true and that all statements made on information or belief are believed to be true; and further that these statements are made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under § 1001 of Title 18 of the United States Code, and that such willful false statements, if made, may jeopardize the validity of the application or any patent issuing thereon.

Pablo Tamayo	1/24/2002 Date
Jill Mesirov	1/24/02 Date
Eric S. Lander	Jan 17, 2002 Date
Todd R. Golub	1/22/02 Date

Self-Organized Maps of Gene Expression: Applications to Hematopoietic Differentiation and the Cell Cycle

Introduction

The recent development of oligonucleotide and cDNA based microarray technologies has enabled investigators to measure the expression levels of large numbers of genes simultaneously (refs). High density arrays produced on nylon membranes, glass slides or silica wafers can now facilitate the monitoring of more than 40,000 genes and ESTs. Many of the technical barriers to complex gene expression measurement appear to have been crossed. It has become apparent, however, that the analytical tools required for such expression analysis are lacking. In particular, automated methodologies for detecting and visualizing gene expression patterns have not yet been reported. Such computational tools will be essential if biologically important insights are to be gleaned from expression profiling experiments. We now describe an automated approach to characterizing gene expression behaviors through the generation of Self-Organized Maps of gene expression.

Self-Organized Maps

The principle behind the Self-Organized Map (SOM) algorithm is that genes with a similar pattern of expression across a series of time points or sample types will be grouped together to form a single cluster of genes. Each group of genes is represented by the centroid (mean) of each cluster; the centroid can therefore be seen as the "signature" of gene expression which is shared by all elements of the cluster. As illustrated in Figure 1A, the gene expression SOM creates a taxonomy of gene expression behaviors whereby clusters exhibiting similar yet distinct behaviors are located adjacent to each other within a two dimensional map. This approach can be compared, for example, to butterfly taxonomy, where butterflies sharing certain characteristics (e.g. size, color) are located adjacent to each other in a specimen tray (Figure 1A). Although the topology of the map (e.g. the number of clusters) is determined by the user, the behaviors to be captured by the map are not pre-determined; rather the behaviors are extracted from the data itself.

The algorithm underlying the gene expression SOM was first described by Kahonen as a model for brain development and competitive learning (ref). The algorithm is a form of neural network which starts with a predefined number of nodes (clusters) connected in a

simple topology such as a two-dimensional grid. Each node is characterized by a weight vector which contains one value for each input variable and is initialized at random. The algorithm sets up an ordering process by iteratively adjusting the node weights so that the nodes migrate to new locations approximating the distribution of input patterns while preserving the topology of the grid. At the end of the process each input pattern is relatively close to one of the nodes. Figure 1B shows an example of a 3x2 SOM topology, its weights, and the ordering process for a simple two-dimensional dataset. The algorithm is conceptually simple, fast, and provides an instant taxonomy and classification scheme of input patterns. For details of the algorithm, refer to our website at http://XXX.

SOM and Hematopoietic Differentiation

To test the SOM approach to classifying gene expression behaviors, we studied the process of hematopoietic differentiation using oligonucleotide expression arrays. We chose hematopoietic differentiation as a model for several reasons: 1) the process is largely controlled at the level of transcription (refs), 2) a block in hematopoietic differentiation is thought to underlie the pathogenesis of leukemia, 3) cell lines which model the process of differentiation are available, and 4) these cell lines have been extensively characterized over the past decade for their expression of individual genes, the majority of which we could now monitor in a single experiment, thereby creating a reference expression database of hematopoietic differentiation.

In the first set of experiments, the myeloid cell line HL-60 was treated with phorbol ester (TPA) which is known to induce macrophage differentiation (refs). Cells were harvested at 0, 0.5, 4 and 24 hours following TPA treatment and polyA+ mRNA was isolated. Double-stranded cDNA was generated from the mRNA, and used for *in vitro* transcription in the presence of biotinylated NTPs, thereby creating biotinylated antisense cRNA. These cRNA targets were hybridized to Affymetrix HU6000 arrays containing 40 25-mer oligonucleotide probes for each of XXX known human genes and XXX ESTs (a complete list of genes present on the arrays as well as the labeling protocol can be obtained through our web site at XXX). Following overnight hybridization and staining with streptavidin-phycoerythrin, the arrays were scanned with an ion argon confocal laser (HP), and the fluorescence intensity of each feature on the array was measured.

As illustrated in Figure 2, following a scaling procedure to account for minor variation in overall chip fluorescence intensities, the intensities for the 40 probes representing each gene

were used to derive a single raw expression level for each gene in each of the 4 samples (for details of these algorithms, see our web site at XXX). A threshold of 20 units was assigned to any gene with a raw expression level less than 20. The next step in generating the SOM was the introduction of a filter which excluded genes with minimal or no variation in expression level across the time course experiments. To be retained in the filtered set, genes were required to change by a factor of at least 3 and have an absolute change in threholded expression level of at least 100 units. Of the 6000 genes on the array, 567 passed the filter. The 567 genes showing variation were then normalized such that the mean expression level within each cell line was 0, and the standard deviation was 1. This normalization process was critical for clustering together genes with similar *patterns* of expression, even if their *absolute* levels of expression differed.

As shown in Figure 3A, the 567 normalized expression levels were used to create a 3X4 (12 cluster) SOM in which distinct expression patterns could be visualized. The entire list of genes contained within each of these clusters can be viewed at our web site, http://XXX. In this approach, each of the 567 genes is represented by one of 12 centroids. For example, cluster 2 contains 64 genes which show down-regulation with time following TPA treatment (Figure 3B). This cluster would be expected to contain genes associated with cell proliferation, since HL-60 cells undergo growth arrest within 24 hours of TPA stimuation. Indeed, included among the elements of Cluster 2 are the genes encoding the cell cycle-related proteins cyclin D2, cyclin D3, CDK2 and PCNA. Similarly, the centroid of Cluster 4 (representing 71 genes whose expression peaks within 30 minutes of TPA treatment) is suggestive of an immediate early response. In support of this, Cluster 4 contains the immediate early genes encoding SRF (serum response factor) and EGR1. Finally, one would expect Cluster 10, corresponding to 142 genes showing late induction to contain genes relating to macrophage differentiation. Cluster 10 accordingly contains the genes encoding the CSF1 receptor, IL1ß and cathepsin B, all known to be involved in macrophage differentiation (refs). Also found within within this cluster are genes not previously known to be regulated in myeloid differentiation, such as the gene encoding the A4 protein reported to be involved in differentiation of colonic epithelium (ref). This experiment demonstrates the ability of the SOM to properly organize a large number of genes into a simple taxonomy of gene expression.

Extending the Complexity of the SOM

To determine whether the SOM approach could be applied to experiments of greater complexity, we extended the HL-60 differentiation experiments to include an additional three hematopoietic cell lines. These cell lines included U937 which, like HL-60, undergoes macrophage differentiation following treatment with TPA, the T-cell line Jurkat which acquires many of the hallmarks of T-cell activation in response to TPA, and the acute promyelocytic leukemia cell line NB4 which undergoes neutrophilic differentiation in response to all trans retinoic acid (ATRA). A total of 17 RNA samples were generated (see *Methods*) which were analyzed on the HU6000 microarrays, thereby generating 102,000 expression levels (6000 genes X 17 samples). Of the 6000 genes assayed, 1063 passed the filter (see above and Figure 2) in at least one of the four cell lines. These 1063 genes in principle could represent 1063 distinct behaviors across the four cell lines. As Figure 4 illustrates, however, the complexity of those 1063 genes could be reduced to a 24-cluster SOM. For example, Cluster 21 represents 21 genes which are upregulated by TPA in the highly related cell lines HL-60 and U937. The adjacent Clusters 17 and 21 reflect genes regulated by TPA in either HL-60 or U937, and Cluster 22 reflects genes regulated in the myeloid cell lines HL-60, U937 and NB4, but not the lymphoid cell line Jurkat.

Cluster 15 of the SOM indicates that 154 genes share a unique pattern of expression, namely upregulation by ATRA in the NB4 cell line and no regulation in the other three cell systems. Of the 154 elements in cluster 15, X encode known markers of neutrophil differentiation (e.g. GCSF receptor, CD59 and Defensin α4). An additional X genes and X ESTs, however, were not previously known to be involved in this process. These genes are of particular biological interest, because they represent clues to the mechanism by which ATRA induces differentiation of acute promyelocytic leukemia cells both in cell culture and in humans with the disease. NB4 cells, like most patients with APL, harbor the PML/RARα fusion protein resulting from a t(15;17) chromosomal translocation which fuses the X protein PML with the retinoic acid receptor α (RARα) (refs).

We chose to further investigate one of these candidate genes, G0S2, encoding a protein of unknown function which was reported as a cyclohexamide inducible protein in human peripheral blood T cells (ref). As detailed in Figure 5, Northern blot analysis confirmed the rapid induction of G0S2 by ATRA in NB4 cells (preceeding morphologic differentiation) and the absence of G0S2 regulation in the other cell lines tested. In addition, G0S2 induction appeared to be specific to cells harboring an intact PML/RARα fusion protein; cells lacking the translocation or NB4 cells harboring a PML/RARα point mutation showed no G0S2 induction. Further studies will be required to elucidate the functional role of

G0S2 in ATRA-mediated differentiation in APL, but the experiments described here illustrate the ability of the SOM to serve as a starting point for the rapid identification of candidate genes worthy of more detailed, functional analysis.

SOM and the Cell Cycle

As a second test of the validity of the SOM approach to gene expression, we investigated the yeast mitotic cell cycle. Cho et al recently reported the identification of *S. cerevisiae* genes which exhibit periodicity restricted to particular phases of the cell cycle (ref). This was accomplished by synchronizing cells in G1 using a temperature-sensitive cdc28-13 allele, releasing the cells from G1 by switching to the non-permissive temperature, and then collecting cells for RNA extraction every ten minutes for a total of 160 minutes (two cell cycles). The expression levels of 6,218 yeast ORFs were measured at each of the 16 time points using Affymetrix oligonucleotide arrays. The 1348 genes which showed greater than 2-fold variation anywhere across the 16 time points were then examined for possible periodicity. This was done by visual inspection of the 1348 genes for reproducible peaks of expression in either the early G1, late G1, S, G2 or M phases of the cell cycle. A total of 416 genes were judged to show such periodicity (ref). Using the raw expression data made publicly available through the investigators' web site (http://genomics.stanford.edu), we sought to determine whether the described periodicity of gene expression could be exposed using a SOM, with no *a priori* knowledge of what patterns the raw data would reflect.

Figure 6 shows the 6x5 (30 cluster) SOM of 828 genes which passed an initial filter (fold change of at least 2 and an absolute change of expression of at least 35 units in *each* of the two complete cell cycles). The SOM map dramatically reduces the number of patterns that require visual inspection from 828 genes to only 30 centroids representing those 828 genes. Examination of the SOM revealed that many of the clusters exhibit periodic behavior. For example, cluster 29 contains 76 genes with peaks in gene expression corresponding to the late G1 phase of the cell cycle (minutes 25-45 and 85-105, as reported by Cho et al (ref)). Also note that the clusters adjacent to cluster 29 (clusters 28 and 24) appear highly related, and also contain genes with late G1 periodicity. The centroids of the SOM-derived clusters corresponding to the G1, S, G2 and M phases of the cell cycle are shown in Figure 7A. The centroids of the previously reported visually derived clusters are shown in Figure 7B. The comparisons of Figures 7A and 7B indicate that the waves of gene expression detected by visual inspection could be rapidly and automatically produced

by the SOM approach. It is noteworthy that the SOM identified these patterns of gene expression independent of any insight into their biological significance.

To confirm that the SOM clustering approach was at least as effective as clustering by visual inspection, we examined the promoters of the genes exhibiting periodicity. We reasoned that an automated clustering approach should retain (if not improve) the previously reported correlation between cell cycle-specific periodicity and particular promoter sequences (refs). We focused on late G1- specific genes because both the visualization- and SOM-generated approaches identified a large number of genes peaking in late G1. As shown in Table 1, Cho et al reported 134 late G1 genes; cluster 29 of our SOM identified 76 genes, and the adjacent clusters 28 and 24 contain an additional 26 and 37 G1-specific genes, respectively, totalling 139 genes. Both the visually clustered and SOM clustered lists of G1 genes were subjected to a search for hexanucleotide sequences within the 500 base pairs upstream of the start codon using an algorithm accessible through http://copan.cifn.unam.mx/Computational_Biology/yeast-tools/. Hexameric sequences which were highly over-represented relative to that expected for the entire genome are shown in Figure 8. The SOM-generated clusters showed a similar, or in some cases, increased ratio of observed-to-expected occurrences of hexameric sequences when compared to visual clustering. These experiments indicate that the SOM was successful in automatically exposing the periodic behavior of cell cycle-related genes.

Conclusions

The experiments presented in this report demonstrate that the Self-Organized Map approach to complex gene expression profiling is a rapid, effective means of classifying microarray expression data. While the experimental data presented here used oligonucleotide microarrays to generate the raw expression values, the analytical methods are equally applicable to other approaches to highly parallel gene expression profiling. The SOM is particularly useful as an initial global view of expression patterns, helping to identify groups of genes deserving closer scrutiny. In the case of the hematopoietic differentiation experiments, we are able to readily focus our attention on a small group of genes which exhibited specific behaviors. In particular, we identified G0S2 as a candidate gene which is not broadly regulated in hematopoietic differentiation, but rather is specifically regulated in acute promyleocytic leukemia cells treated with retinoic acid. Similarly, the SOM methodology exposed the periodicity of cell cycle related genes without the need to visually inspect each gene's individual pattern of expression. Such an automated approach will be

particularly important when microarrays containing the entire human genome become available; visual inspection of 100,000 genes will not be feasible. Perhaps most significantly, however, the SOM approach has the potential to expose gene expression patterns that were not previously anticipated. In that regard, the SOM represents an exploratory tool which functions as an organizational framework created by the data itself, not by a preconceived biological hypothesis. We believe that such analytical tools are a necessary first step in extending the value of expression profiling beyond our current understanding of coordinate gene regulation. *(2647 words)*

Figures and Tables

Fig 1: Principle of the SOM

A) schematic of SOM clustering and butterfly comparison
B) schema of SOM iteration

Fig 2: Data Processing Schema

scan-->intensities-->scaling-->GEprocess-->thresholding-->
filtering-->normalization-->clustering-->display centroids

Fig 3: Hematopoietic SOM

A) 3X4 SOM for HL60 TPA differentiation
B) detail of cluster 2 (67 down-regulated genes)

Fig 4: 6X4 multipanel hematopoietic differentiation SOM

Fig 5: G0S2 regulation

A) chip values for G0S2 in NB4 cells plus ATRA (time course)
B) Northern blot confirmation
C) NB4 time course photomicrographs
D) Northerns of: NB4+DMSO; HL60+ATRA; HL60+DMSO
E) Northern of ATRA-resistant NB4 cells

Fig 6: Yeast mitosis SOM

Fig 7:

A) SOM-generated periodic centroids
B) visually-generated periodic centroids

Fig 8: Promoter element comparison plot of our vs their G1-phase clusters

Table 1: Elements of visual and our G1-phase clusters

Fig 1A

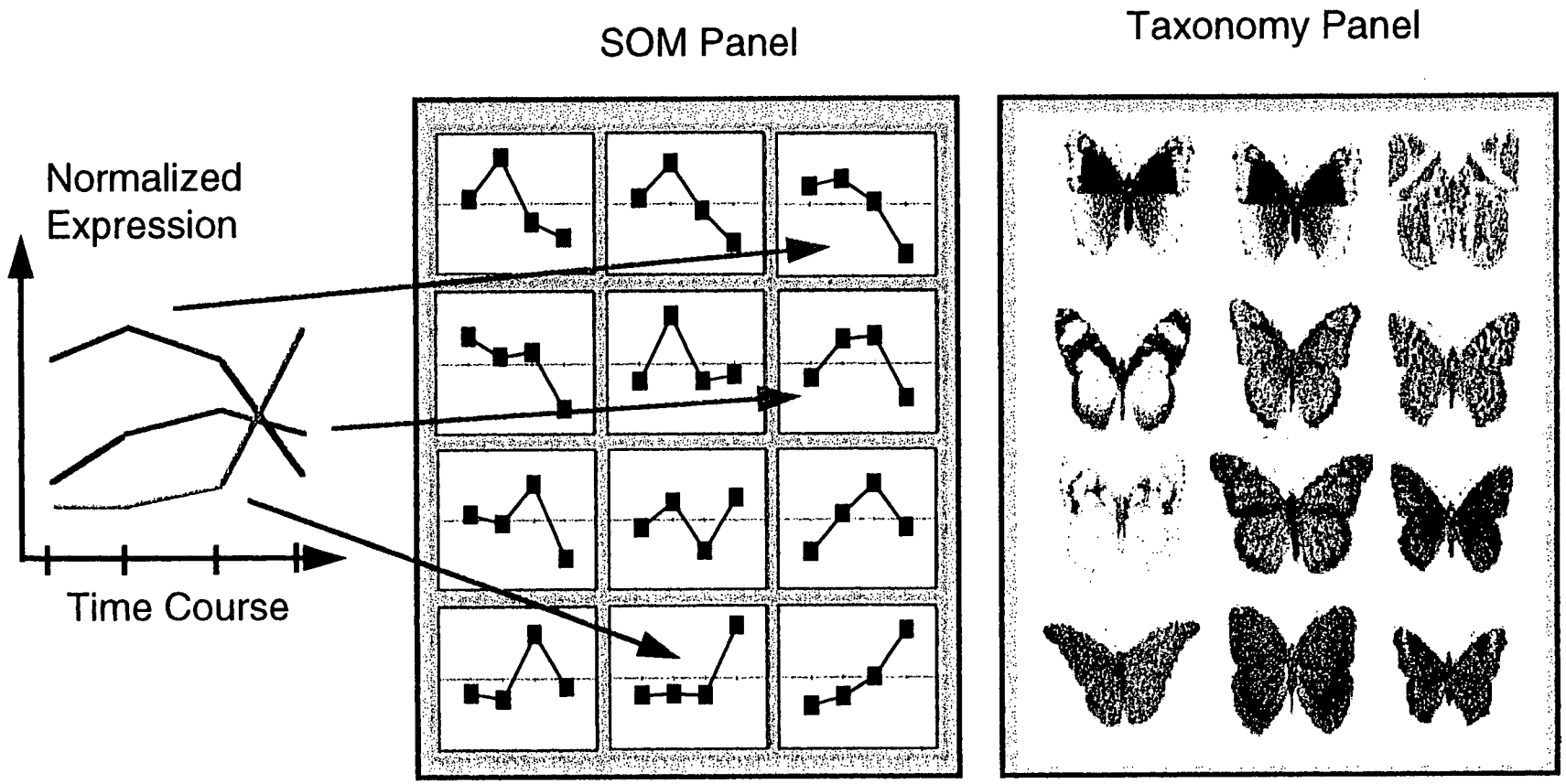
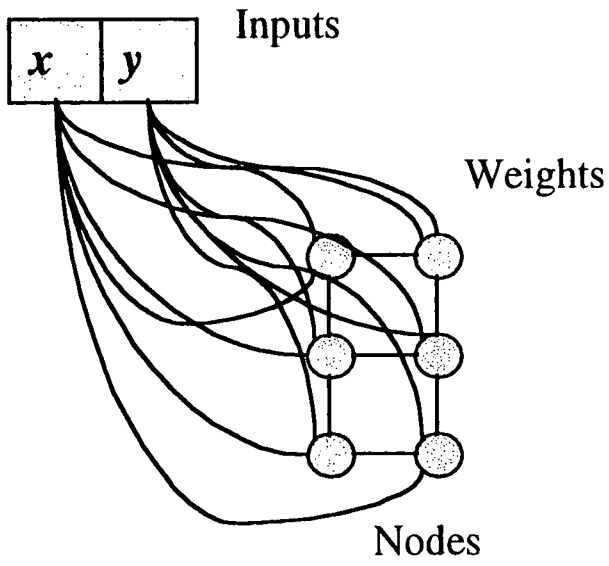

SOM Panel
Taxonomy Panel
Normalized Expression
Time Course

Fig 1B

3 x 2 SOM Rectangular Map



SOM Ordering Process

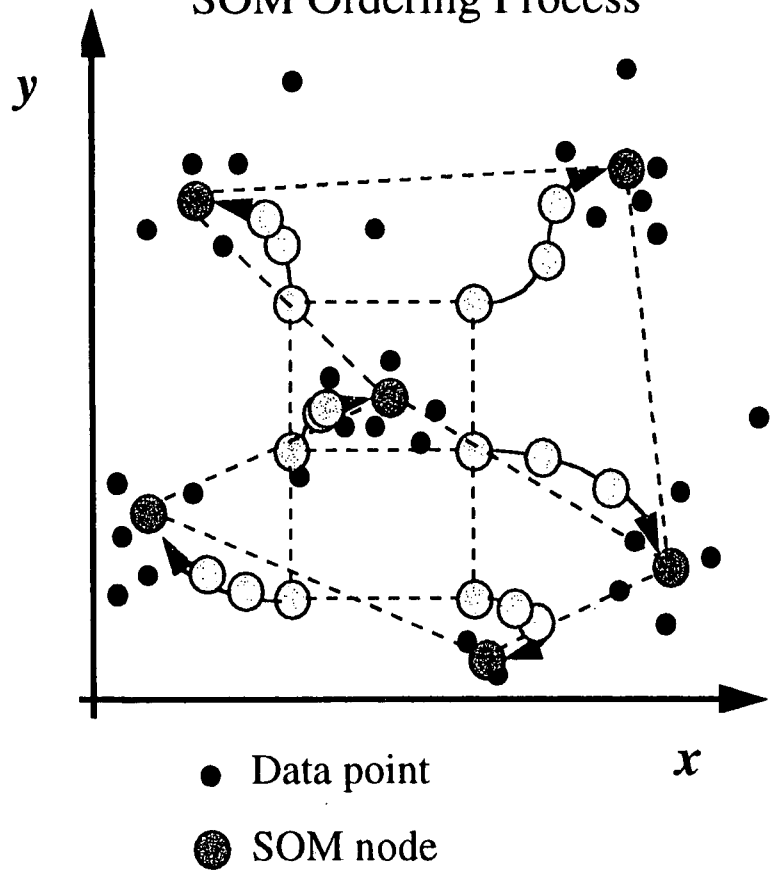


Fig 2.

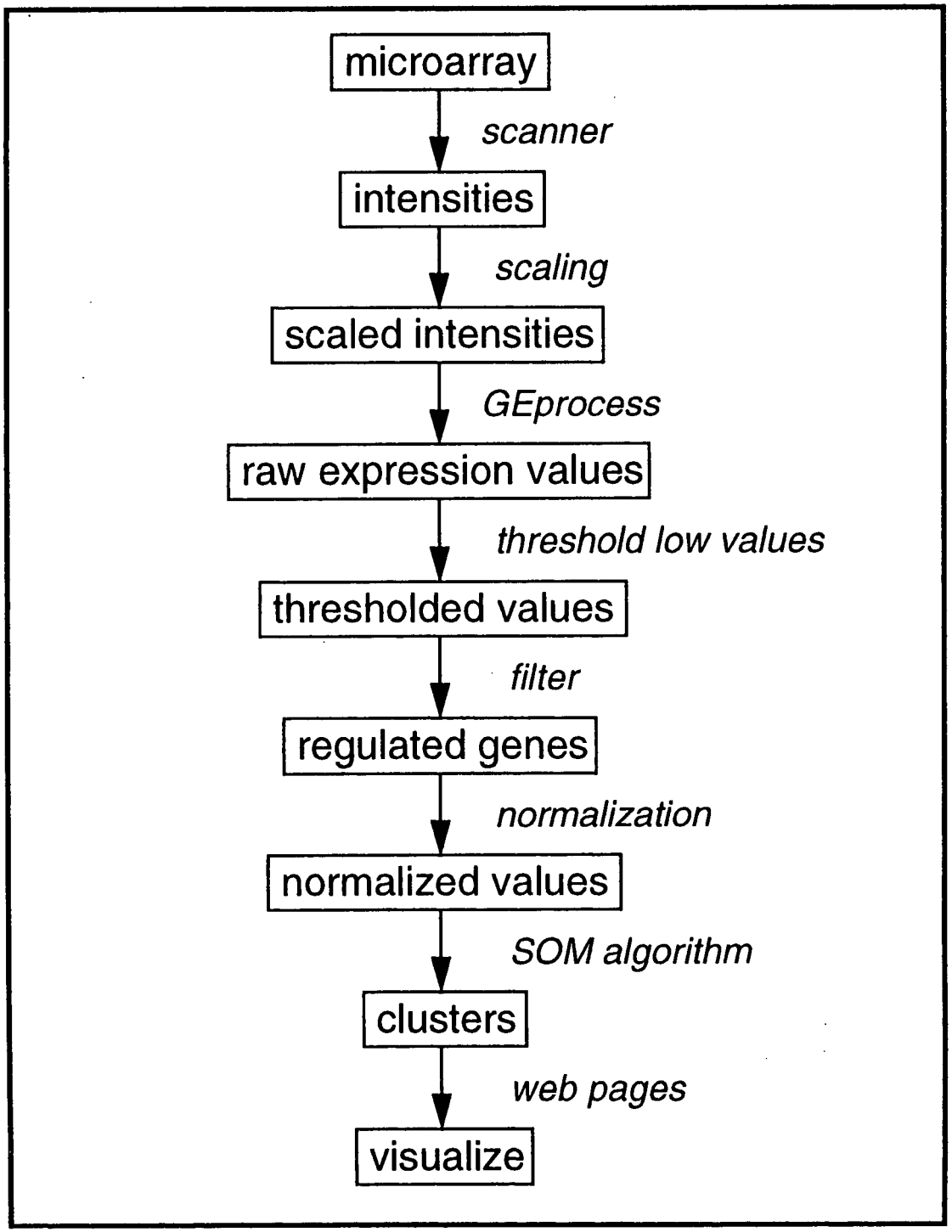
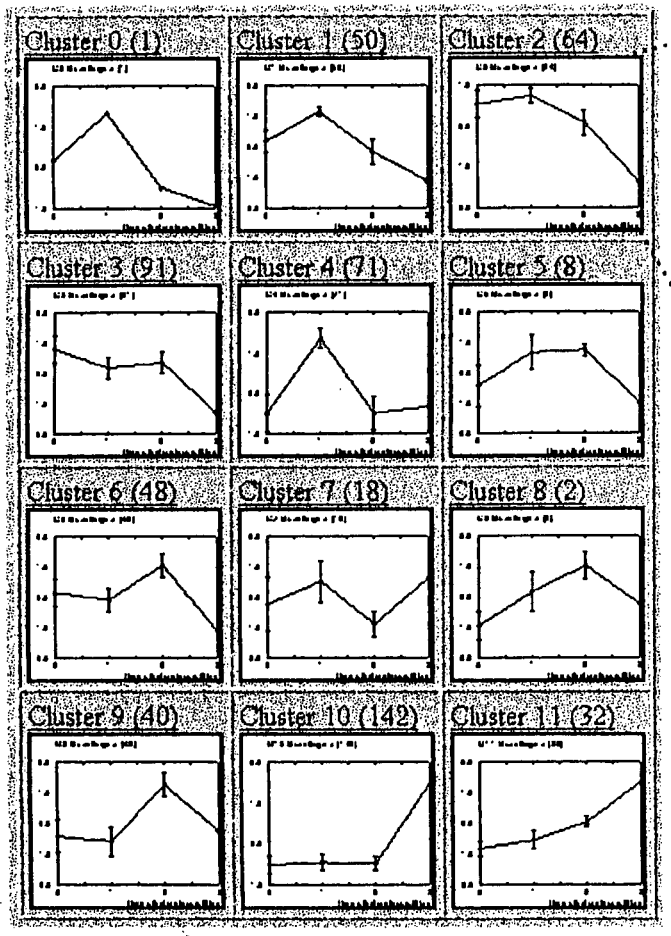

microarray
scanner
intensities
scaling
scaled intensities
GEprocess
raw expression values
threshold low values
thresholded values
filter
regulated genes
normalization
normalized values
SOM algorithm
clusters
web pages
visualize

Fig 3

HL60 SOM Map



Sample Elements (cluster 2)

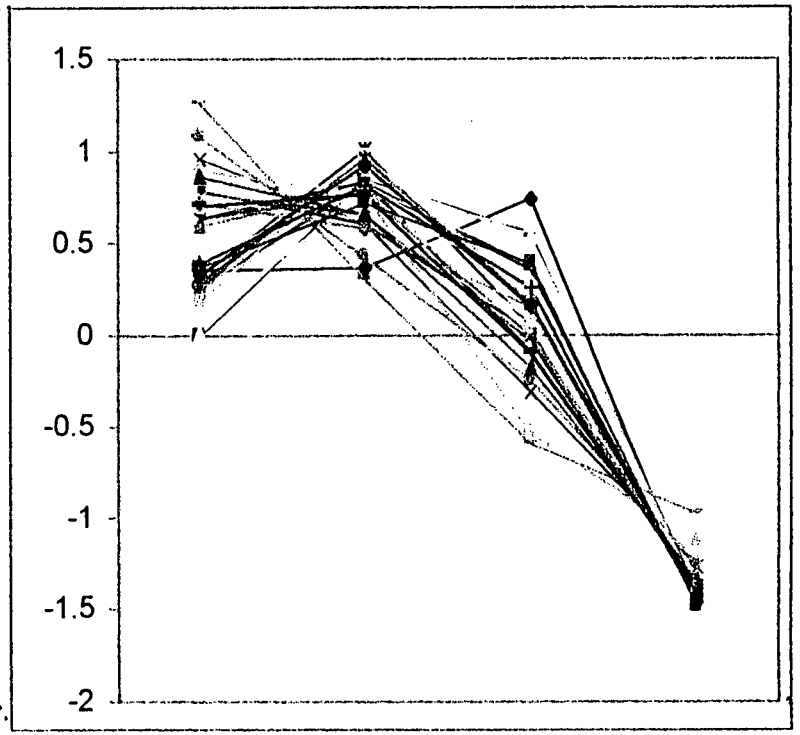


Fig 4

Multi Panel SOM Map: HL60, U937, NB4, Jurkat

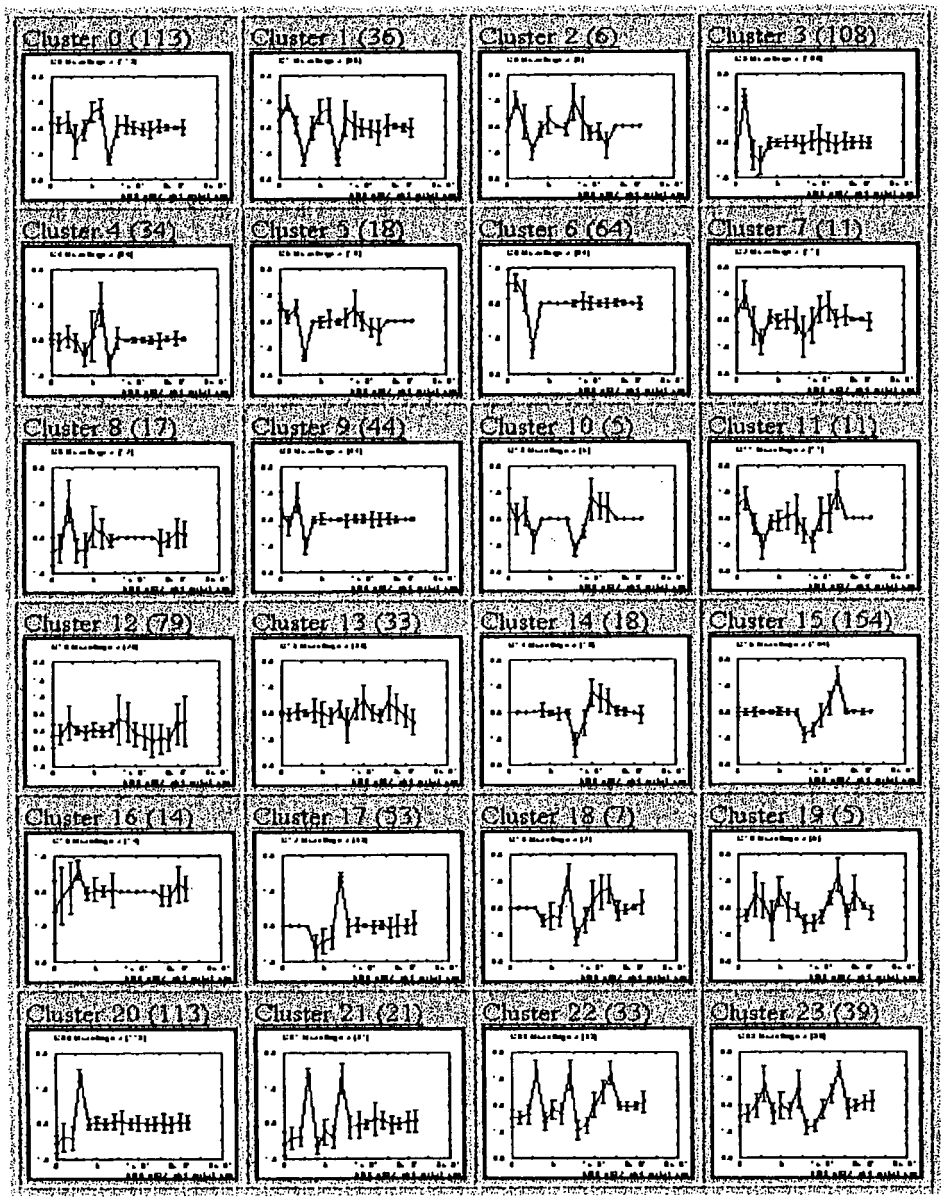
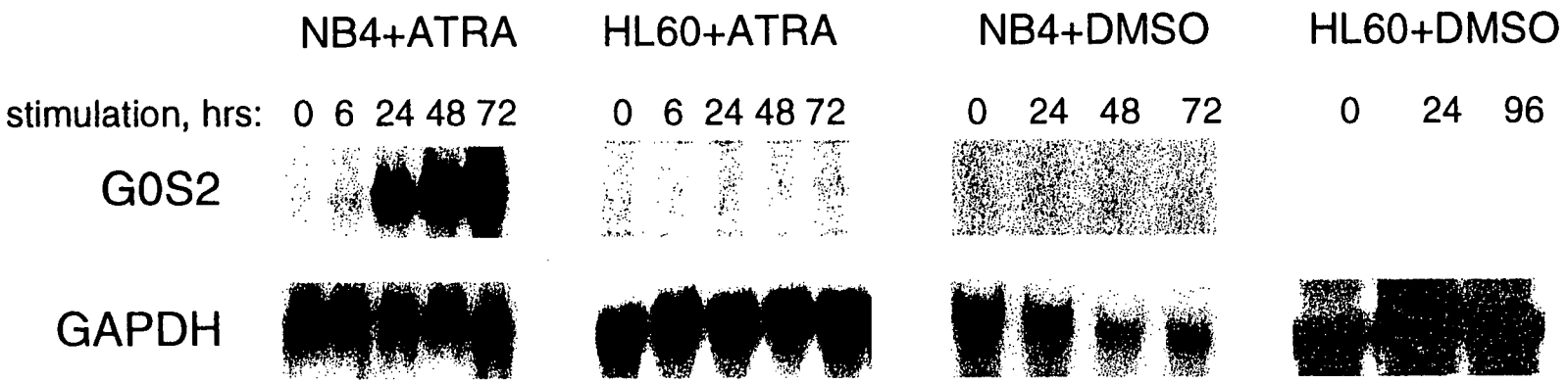


Fig5

Chip Expression Values for NB4+ATRA

stimuation, hrs	0	6	24	48	72
G0S2	4	73	1959	2456	2404
GAPDH	2550	2341	2300	2615	2148



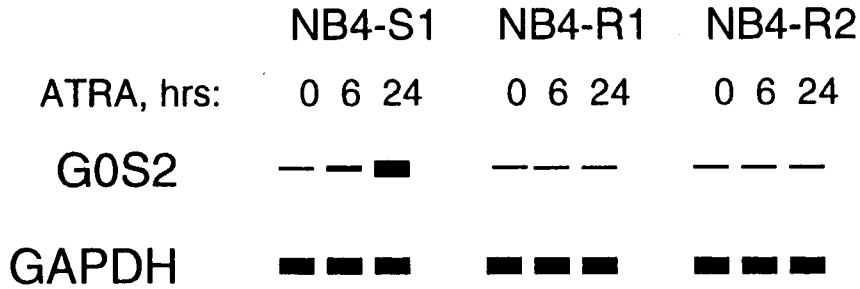


Fig 6

Yeast Cell Cycle SOM Map

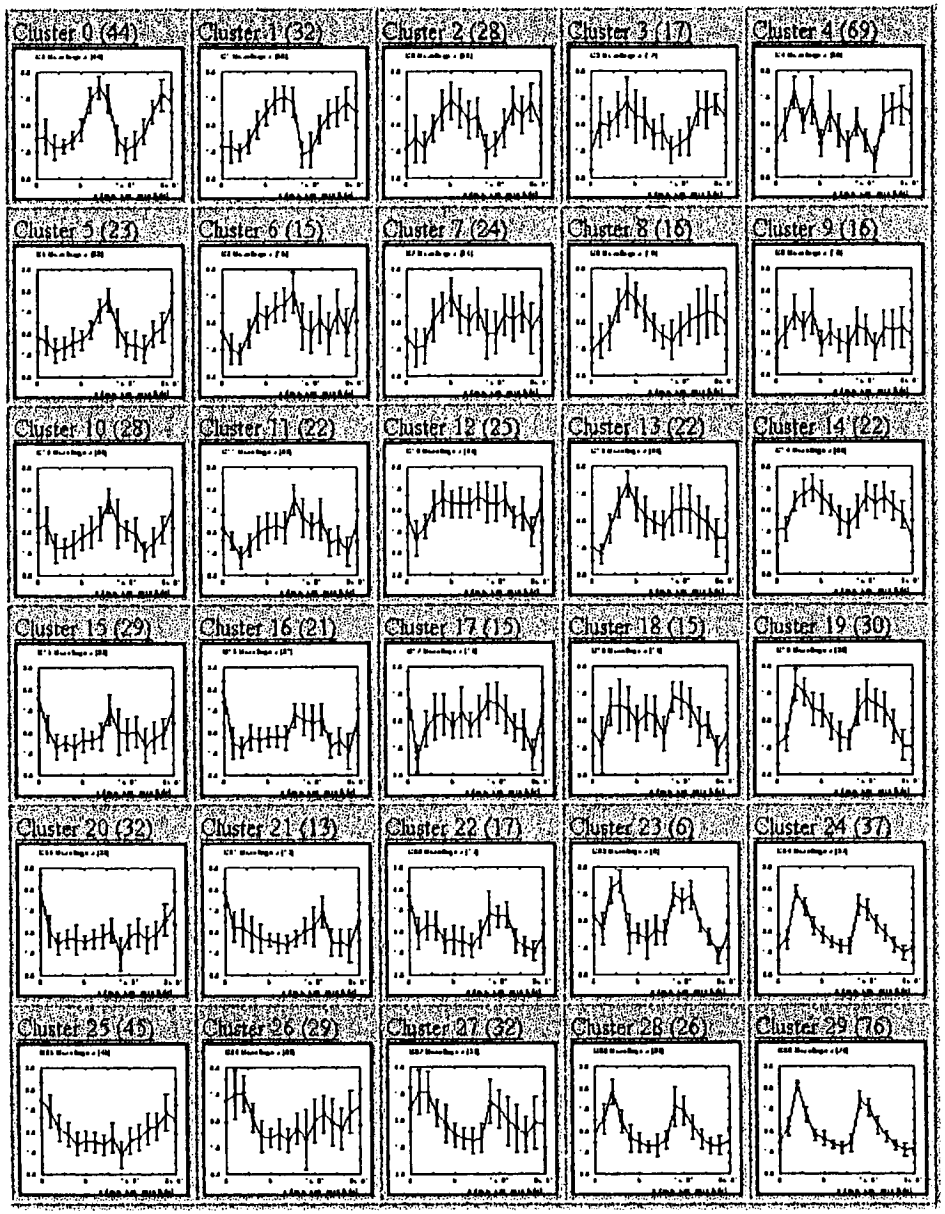
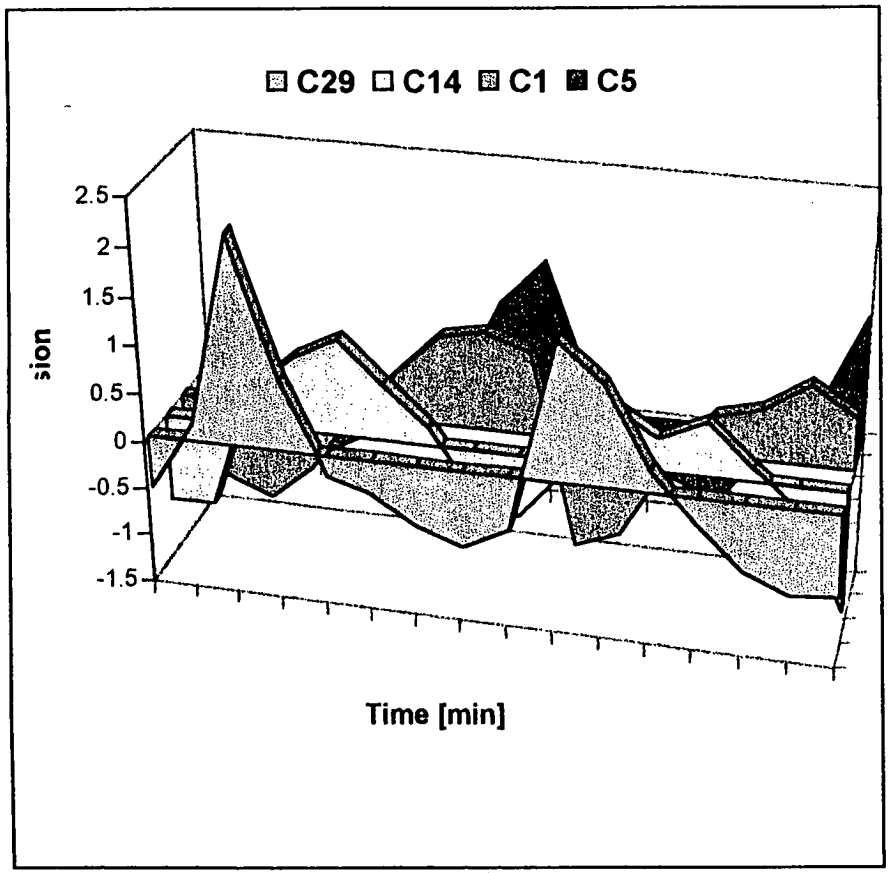


Fig 7


C29 C14 C1 C5
sion
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
Time [min]

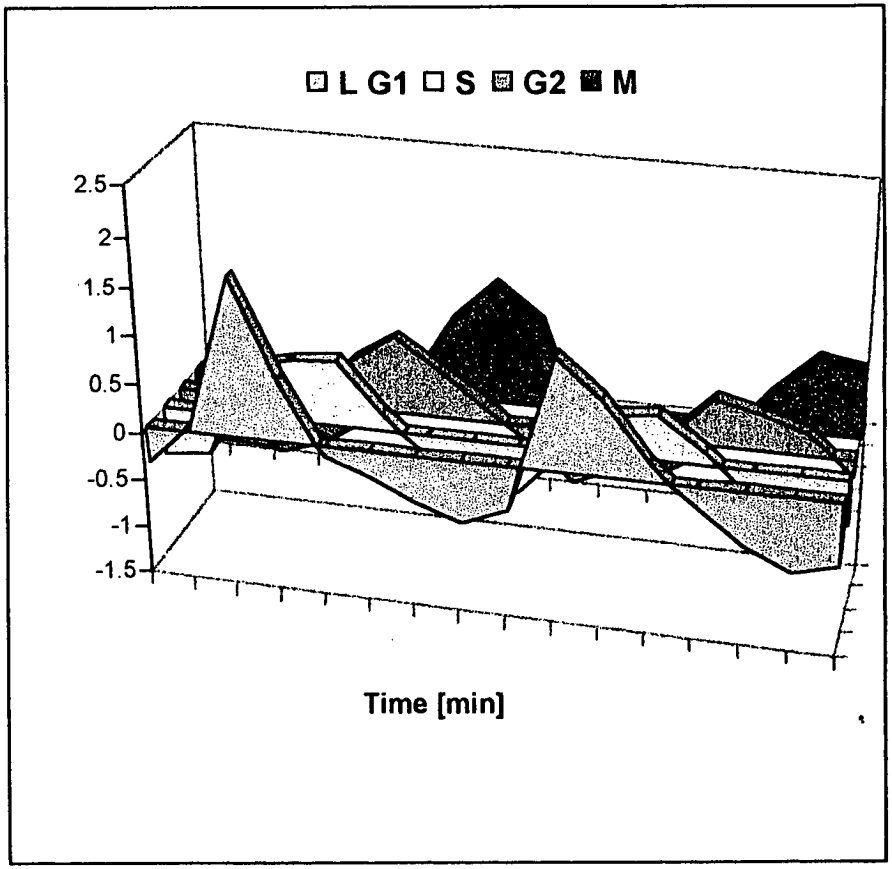
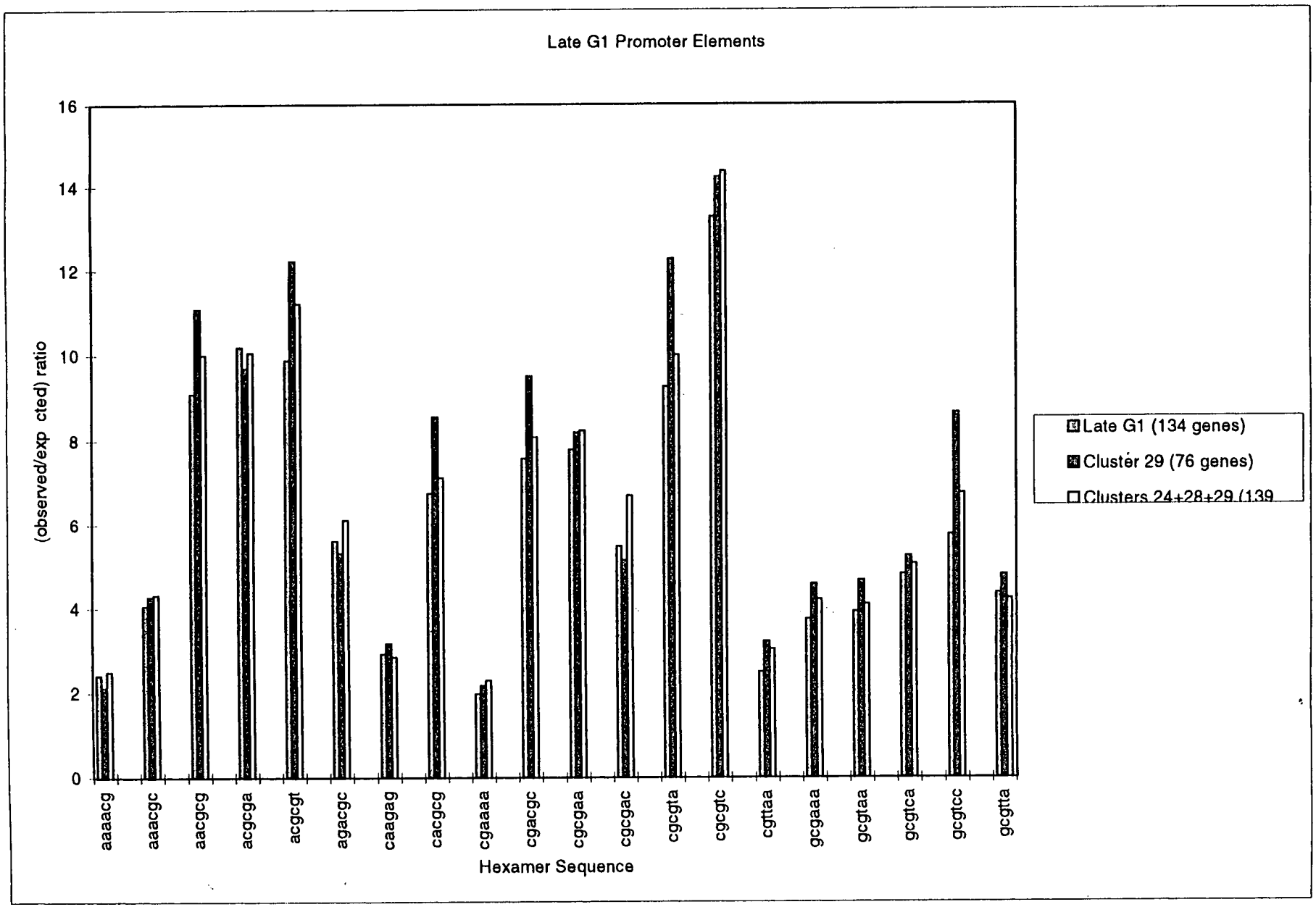

L G1 S G2 M
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
Time [min]

Fig8.


Late G1 Promoter Elements
(observed/exp cted) ratio
16
14
12
10
8
6
4
2
0
aaaacg
aaacgc
aacgcg
acgcga
acgcgt
agacgc
caagag
cacgcg
cgaaaa
cgacgc
cgcgaa
cgcgac
cgcgta
cgcgtc
cgttaa
gcgaaa
gcgtaa
gcgtca
gcgtcc
gcgtta
Hexamer Sequence
Late G1 (134 genes)
Cluster 29 (76 genes)
Clusters 24+28+29 (139

EXHIBIT

B

WHITEHEAD INSTITUTE

DISCLOSURE NOTICE

To: Elliott Sigal, Bill Koster, Marilyn Hartig, Joseph Sorrentino, Charles Linzner (BMS)
Tom Gingeras, Rob Lipshutz, Philip McGarrigle, Steve Fodor (Affymetrix)
Bob Tepper, Frank Lee, Keith Dionne, Steve Holtzman (MPI)

From: Gwen Acton, Functional Genomics Program Manager

Date: October 23, 1998

RE: Self-Organized Maps of Gene Expression

cc: Eric Lander, John Pratt, Gerry Fink (WI), Pat Granahan (HBSR)

In accordance [redacted] we are providing to Consortium Members 30 day Notice prior to submission for publication or disclosure of data and information developed at Whitehead as part of the Functional Genomics Program.

Enclosed please find a copy of a draft of a paper entitled *"Self-Organized Maps of Gene Expression: Applications to Hematopoietic Differentiation and the Cell Cycle"* that the Molecular Pattern Recognition group of the Functional Genomics Program plans to submit for publication to the journal Nature Genetics.

The paper describes an automated approach to characterizing gene expression data using Self-Organized Maps (SOM) which cluster genes with similar expression patterns. SOM was tested by comparing gene expression levels in cells in different stages of hematopoietic differentiation. Using this methodology, the group identified GOS2 as a candidate gene which is specifically regulated in acute promyleocytic leukemia cells treated with retinoic acid. In a separate experiment, the SOM methodology revealed the periodicity of cell cycle related genes in yeast without need of visual inspection.

The results suggest that self-organized maps are a rapid and efficient means of classifying microarray expression data. SOM methodology appears to be particularly useful as initial global analysis of expression patterns, helping to identify groups of genes for closer scrutiny.

[redacted]

Whitehead Institute / MIT Center for Genome Research / One Kendall Square, Building. 300, Cambridge, MA 02139-1561
617/252-1900 FAX 617/252-1902